Illuminating the dynamic rare biosphere of the Greenland Ice Sheet's Dark Zone

Gokul, Jarishma K; Cameron, Karen A; Irvine-Fynn, Tristram D L; Cook, Joseph M; Hubbard, Alun; Stibal, Marek; Hegarty, Matt; Mur, Luis A J; Edwards, Arwyn

Published in:
FEMS Microbiology Ecology

DOI:
10.1093/femsec/fiz177

Publication date:
2019

Citation for published version (APA):
Gokul, J. K., Cameron, K. A., Irvine-Fynn, T. D. L., Cook, J. M., Hubbard, A., Stibal, M., Hegarty, M., Mur, L. A. J., & Edwards, A. (2019). Illuminating the dynamic rare biosphere of the Greenland Ice Sheet's Dark Zone. *FEMS Microbiology Ecology*, *95*(12), [fiz177]. https://doi.org/10.1093/femsec/fiz177

Unlinked rRNA genes are widespread among Bacteria and Archaea

Authors: Tess E. Brewer[a,b,*], Mads Albertsen[c], Arwyn Edwards[d], Rasmus H. Kirkegaard[c], Eduardo P. C. Rocha[e], Noah Fierer[a,f]

[a] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309 USA
[b] Current Address: Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland
[c] Department of Chemistry and Bioscience, Aalborg University, 9220 Aalborg, Denmark
[d] Institute of Biological, Environmental and Rural Sciences, Aberystwyth University SY23 3DA UK
[e] Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris, 75015, France.
[f] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309 USA

* Corresponding author:
Tess Brewer
University of Zürich
Building Y27-J-54
Winterthurerstrasse 190
Zürich, Switzerland 8057
Email: tess.brewer@uzh.ch

Abstract

Ribosomes are essential to cellular life and the genes for their RNA components are the most conserved and transcribed genes in Bacteria and Archaea. Ribosomal rRNA genes are typically organized into a single operon, an arrangement thought to facilitate gene regulation. In reality, some Bacteria and Archaea do not share this canonical rRNA arrangement - their 16S and 23S rRNA genes are separated across the genome and referred to as "unlinked". This rearrangement has previously been treated as an anomaly or a byproduct of genome degradation in intracellular bacteria. Here, we leverage complete genome and long-read metagenomic data to show that unlinked 16S and 23S rRNA genes are more common than previously thought. Unlinked rRNA genes occur in many phyla, most significantly within Deinococcus-Thermus, Chloroflexi, and Planctomycetes, and occur in differential frequencies across natural environments. We found that up to 41% of

rRNA genes in soil were unlinked, in contrast to the human gut, where all sequenced rRNA genes were linked. The frequency of unlinked rRNA genes may reflect meaningful life history traits, as they tend to be associated with a mix of slow-growing free-living species and intracellular species. We speculate that unlinked rRNA genes may confer selective advantages in some environments, though the specific nature of these advantages remains undetermined and worthy of further investigation. More generally, the prevalence of unlinked rRNA genes in poorly-studied taxa serves as a reminder that paradigms derived from model organisms do not necessarily extend to the broader diversity of Bacteria and Archaea.

Introduction

Ribosomes are the archetypal "essential proteins", so much so that they are a key criteria in the division between cellular and viral life (1). In Bacteria and Archaea, the genes encoding the RNA components of the ribosome are traditionally arranged in a single operon in the order 16S - 23S - 5S. The rRNA operon is transcribed into a single RNA precursor called the pre-rRNA 30S, which is separated and processed by a number of RNases (2). This arrangement of rRNA genes within a single operon is thought to allow rapid responses to changing growth conditions - the production of rRNA under a single promoter allows consistent regulation and conservation of stoichiometry between all three, essential components (3). Indeed, the production of rRNA is the rate-limiting step of ribosome synthesis (4), and fast-growing Bacteria and Archaea accelerate ribosome synthesis by encoding multiple rRNA operons (5).

Some Bacteria and Archaea have "unlinked" rRNA genes, where the 16S and 23S rRNA genes are separated by large swaths of genomic space (Figure 1). This unlinked rRNA gene arrangement was first discovered in the thermophilic bacterium *Thermus thermophilus* (6). Reports of unlinked rRNA genes soon followed in additional Bacteria, including the planctomycete *Pirellula marina* (7), the aphid endosymbiont *Buchnera aphidicola* (8), and the intracellular pathogen *Rickettsia prowazekii* (9). Though unlinked rRNA genes were first discovered in a free-living environmental bacterium, their ubiquity among the order Rickettsiales has led to suggestions that unlinked rRNA genes are a result of the genome degradation typical of obligate intracellular lifestyles (10-12).

With this study we sought to determine the frequency of unlinked rRNA genes across Bacteria and Archaea and whether this unique genomic feature is largely confined to those Bacteria and Archaea with an obligate intracellular lifestyle. We examined the rRNA genes of over 10,000 publicly available complete bacterial and archaeal genomes to identify which taxa have unlinked rRNA genes and to determine if there are any genomic characteristics shared across taxa with this feature. As complete genomes are not typically available for the broader diversity of Bacteria and Archaea found in environmental samples (13), we also characterized rRNA gene arrangements using long-read metagenomic datasets obtained from a range of environmental samples, which together encompassed over 17 million sequences (≥1000 bp). With these long-read metagenomic datasets, we were able to determine whether unlinked rRNA genes are common in environmental populations and how the distributions of unlinked rRNA genes differ across prokaryotic lineages and across distinct microbial habitats.

Methods

Analyses of complete genomes

We downloaded all bacterial and archaeal genomes in the RefSeq genome database (14) classified with the assembly level "Complete Genome" from NCBI in January 2019 (12539 genomes). We removed genomes from consideration that had rRNA genes that were split across the genome start and end (96 genomes), >20 reported rRNA genes (2 genomes), or an unequal number of 16S and 23S rRNA genes (219 genomes). This left us with a set of 12222 genomes. We used gene ranges associated with each open reading frame (ORF) to pair the 16S and 23S rRNA genes that were closest to each other in each genome. We then checked for gene directionality (sense/antisense) and calculated the distance between each pair, taking directionality into account (see Supplemental Figure S1 for more detail and a visual representation). rRNA pairs were classified as 'unlinked' if the distance between each gene was greater than 1500 bp, 'linked' if the distance was less than or equal to 1500 bp. We separated genomes that had a 16S or 23S rRNA gene that started or ended within 1500 bp of the beginning or end of its genome and classified these 226 genomes independently to account for the circular nature of bacterial and archaeal genomes. For this subset of genomes, we iteratively adjusted the start and end position of

those “edge-case” rRNA genes with respect to genome size and selected the smallest distance between the 16S and 23S rRNA genes as the true distance, using the same formula presented in Supplemental Figure 1. Each genome was classified as 'unlinked', 'linked', or 'mixed' depending on the status of their rRNA genes with 'mixed' genomes having multiple rRNA copies with a combination of linked and unlinked rRNA genes. We re-assigned taxonomy to each genome using the SILVA 132 SSU database (clustered at 99%) to maintain a consistent taxonomy between our two datasets. All analyses were done in R version 3.5.1 (15). Information on all genomes included in these analyses (including classification of rRNA genes) is available in Supplemental Dataset S1.

Long-read metagenomic analyses

To investigate the prevalence of unlinked rRNA genes among those Bacteria and Archaea found in environmental samples (including many taxa for which genomes are not yet available), we analyzed long-read metagenomic datasets generated from soil, sediment, activated sludge, anaerobic digesters, and human gut samples. These metagenomic datasets were generated using either the Oxford Nanopore MinION/PromethION (6 samples) or the Illumina synthetic long-read sequencing technology (also known as Moleculo, first described in (16), 9 samples). The Moleculo sequences originated from four previously published studies covering: the human gut (17), prairie soil (18), sediment (19), and grassland soils (MG-RAST project mgp14596, (20). The Nanopore sequences originated from four unpublished studies spanning a diverse range of environment types: anaerobic digesters, activated sludge, sediment, and lawn soil. For these samples, DNA was extracted using DNeasy PowerSoil Kits (Qiagen, DE) and libraries were prepared for sequencing using the LSK108 kit (Oxford Nanopore Technologies, UK) following the manufacturers protocol. The libraries were sequenced on either the MinION or the PromethION sequencing platforms (Oxford Nanopore Technologies, UK). Base calling was conducted using Albacore v. 2.1.10 for the lawn soil sample (VCsoil) and Albacore v. 2.3.1 for all other samples (Oxford Nanopore Technologies, UK). Across these 15 samples, we compiled 16,870,533 Nanopore sequences and 846,437 Moleculo sequences with a minimum read length of 1000 bp.

We trimmed the first 250 bp of each Nanopore sequence to remove low quality

regions, but performed no other quality filtering as not all samples included information on sequence quality (some sequences were fasta format). Instead, we relied on our downstream filtering steps to remove sequences of poor quality. Metaxa2 version 2.1 (21) was run on all sequences with default settings to search for SSU (16S rRNA) and LSU (23S rRNA) gene fragments. Taxonomy was assigned to the partial rRNA sequences using the RDP classifier (22) and the SILVA 132 SSU and LSU databases (both clustered at 99% sequence identity, 23). If a sequence contained both 16S and 23S rRNA genes we used the taxonomy with the highest resolution (if the 16S was annotated to family level while the 23S was genus level, we used the 23S taxonomy for both rRNAs). Details on each sample, including number of reads and median read lengths, are available in Supplemental Table S1.

We next used a number of criteria to filter the reads included in downstream analyses and to identify taxa with unlinked rRNA genes. We only included those reads in our final dataset that met the following criteria:

1) Contained a 16S rRNA gene (to avoid potentially double counting organisms with unlinked 16S and 23S rRNA genes),
2) Included the last two domains of the 16S rRNA gene (V8|V9) (Metaxa2 uses multiple Hidden Markov Model (HMM) profiles targeting conserved regions of rRNA genes, each of these regions is referred to as a domain),
3) The length of the 16S rRNA gene was ≤4000 bp and the length of the 23S rRNA gene (if present) was ≤ 6800 bp. These thresholds were chosen to remove erroneously long rRNA genes while accommodating insertions within rRNA genes such as those that occur in Candidate Phyla Radiation (CPR) taxa (24), *Nostoc*, *Salmonella*, and others (25),
4) Could be classified to at least the phylum level of taxonomic resolution.

Of the subset of reads that met these criteria (112 - 878 per Moleculo sample, 3817 - 28056 per Nanopore sample, see Supplemental Table S1 for details), we classified reads as containing unlinked rRNA genes if there was >1500 bp between the 16S and 23S rRNA genes, or if there was no 23S domain found 1500 bp after the end of the 16S rRNA. We note that, unlike the NCBI gene ranges, Metaxa2 takes strand information into account and

translates start and stop locations into sense orientation for SSU and LSU. For our final analyses, we removed reads that could not be classified as linked or unlinked rRNA genes (for instance a sequence with only 300 bp after the 3' end of the 16S rRNA gene). All analyses were done in R version 3.5.1 (15). Information on all long-read sequences included in these analyses (including classification of rRNA genes) is available in Supplemental Dataset S2.

Phylogenetic tree combining long-read and NCBI datasets

A phylogenetic tree was created from full-length 16S rRNA gene sequences by combining both the NCBI complete genomes and representatives of the long-read metagenomic datasets. For the NCBI genome sequences, we selected one 16S rRNA gene sequence per unique species. For the long-read datasets, we first matched the partial 16S rRNA genes recovered by metaxa2 (21) to full-length 16S rRNA gene sequences in the SILVA 132 SSU database (23) using the usearch10 version 10.0.240 command usearch_global (settings: -id 0.95 -strand both -maxaccepts 0 -maxrejects 0; 26). The full-length SILVA 16S rRNA genes sequences that matched to the long-read sequences ≥95% percent identity and ≥500 bp alignment length were used as representatives of their long-read sequence match. We used 95% percent identity as our cutoff as we found unlinked rRNA gene status to generally be conserved within genera (see below and Supplemental Figure S2). The NCBI and SILVA sequences were then aligned with PyNAST version 0.1 (27) and the phylogenetic tree was constructed using FastTree version 2.1.10 SSE3 (28), and plotted with iTOL (29).

Genomic attributes associated with unlinked rRNA genes

All tests for genomic attributes were done with a subset of our complete genome dataset - we reduced the dataset to include only one representative genome per unique species and operon status. For example, if a species had 24 genomes with linked rRNA genes and 3 genomes with unlinked rRNA genes, we retained two genomes total, one linked and one unlinked. Species with heterogeneous rRNA gene status accounted for only 0.71% of species and we found that the presence of unlinked rRNA genes was strongly conserved at the species and genus level (Supplemental Figure S2).

With this set of reduced genomes (3967 genomes in total), we first calculated Pagel's lambda (30) to determine whether there was a phylogenetic signal associated with unlinked rRNA genes using the phylosig function of the phytools package version 0.6.60 (31). The results of this test indicated there was a strong phylogenetic signal (lambda = 0.96, $p < 0.0001$), so we controlled for phylogeny in all of our subsequent tests by using a Phylogenetic Generalized Linear Model for continuous variables (with the function phyloglm in the phylolm package version 2.6; 32).

To determine if taxa with unlinked rRNA genes have a lower predicted growth rate, we calculated the codon usage proxy ΔENC' (33,34), which provides an estimate of minimum generation times (35). We calculated ΔENC' with the program ENCprime (33) with default options, on both the concatenated ORF sequences and concatenated ribosomal protein sequences for each genome following Vieira-Silva and Rocha (2009). To determine if RNaseIII was present in each genome, we used HMMER version 3.1b2 (36) to search for three RNaseIII pfams (bacterial PF00636, PF14622, and archaeal PF11469) in the translated protein files of each genome. We used the gathering thresholds (GA) associated with each of these pfams to set all cutoffs and reduce the likelihood of false positives (--cut_ga).

Results

Unlinked rRNA genes occur frequently in complete genomes

We used a set of 12222 "complete" bacterial and archaeal genomes extracted from NCBI in January 2019 to determine how frequently unlinked 16S and 23S rRNA genes occur. We analyzed the distribution of distances between the closest edges of the closest pairs of 16S and 23S rRNA genes (known as the Internally Transcribed Spacer - ITS) in each genome and found that the vast majority of 16S and 23S rRNA gene pairs (98.7%) had an ITS $\leq$ 1500 bp with an average ITS length of 418.7 bp (±169.7 bp, Figure 2A). However, pairs with ITS lengths $>$ 1500 bp showed a scattered distribution of distances, with an average ITS length of 410374 bp (±521792 bp). Hence, for this classification scheme we called rRNA genes "unlinked" if the ITS was greater than 1500 bp in length. This 1500 bp threshold is in some ways conservative, as the distance between genes in an operon is

usually quite low - peaking between 20 and 30 bp in most genomes (37). Additionally, tRNA are the most common genes found in the space between the 16S and 23S rRNA genes, and range from only 75 to 90 bp in length (38).

After classifying each rRNA gene pair as linked or unlinked based on the distance between the 16S and 23S rRNA genes, we found that 3.65% of the genomes in our dataset had exclusively unlinked rRNA genes, 0.62% had mixed rRNA gene status (i.e. genomes with multiple rRNA copies that had at least one set of unlinked rRNA genes and at least one canonical, linked rRNA operon), and 95.73% had exclusively linked operons (these numbers do not match up with the per rRNA gene dataset as each genome has a variable rRNA copy number). We found unlinked genomes to be relatively common (present in ≥5% of members) in taxa characterized as having an obligate intracellular lifestyle within the phyla Spirochaetes (genus *Borrelia*), Epsilonbacteraeota (family Helicobacteraceae), Alphaproteobacteria (order Rickettsiales), and Tenericutes (species *Mycoplasma gallisepticum*). However, we also found high proportions of unlinked rRNA genes in phyla that are generally considered to be free-living, such as Deinococcus-Thermus (families Thermaceae and Deinococcaceae), Chloroflexi (family Dehalococcoidaceae), Planctomycetes (families Phycisphaeraceae and Planctomycetaceae), and Euryarchaeota (class Thermoplasmata). Phyla with at least 5% of genomes having exclusively unlinked rRNA genes are shown in Figure 2C.

Unlinked rRNA genes are widespread in environmental metagenomic data

While the results from our complete genome dataset demonstrate that unlinked rRNA genes are common in some putatively free-living phyla, databases featuring complete genomes do not capture the full breadth of microbial diversity and are heavily biased towards cultivated organisms relevant to human health (13). Just three phyla (Proteobacteria, Firmicutes, Actinobacteria) accounted for >83% of the genomes in our NCBI dataset - even though recent estimates of bacterial diversity total at least 99 unique phyla (39). To investigate the ubiquity of unlinked rRNA genes among those taxa underrepresented in 'complete' genome databases, we analyzed long-read metagenomic data from a range of distinct sample types. Focusing exclusively on long-read sequences

allowed us to span the 1500 bp distance required for classification of rRNA genes without the need for assembly. This is important as the repetitive structure of rRNA genes makes it difficult to assemble a mix of non-identical rRNA genes from the short reads typical of most current metagenomic sequencing projects (40).

From our initial long-read dataset encompassing 15 unique samples (~890,000 Illumina synthetic long reads (also known as Moleculo) and ~19 million Nanopore reads, with median read lengths of 8858 bp and 5398 bp, respectively), only 15855 sequences contained rRNA genes and met the criteria we established for the classification of rRNA genes as linked or unlinked (see Methods). Of these reads, we classified 1607 as unlinked, or 10.1% of the dataset (Figure 2B). These long-read metagenomic analyses showed that unlinked rRNA genes are not equally distributed across environments - we found that up to 41% of the taxa in soil had unlinked rRNA genes, whereas other environments had much lower proportions, most notably the human gut, where all sequenced rRNA genes were linked (Figure 3).

The results from our analyses of the long-read dataset generally mirrored the corresponding results from the complete genome dataset, in that many of the long reads classified as unlinked belonged to the same phyla where unlinked rRNA genes were prevalent in the complete genome dataset (Figure 2). The long-read metagenomic dataset confirmed that members of the phyla Deinococcus-Thermus, Planctomycetes, Chloroflexi, Spirochaetes, and Euryarchaeota frequently have unlinked rRNA genes (Figure 2B). The long-read dataset also allowed us to provide additional evidence for unlinked rRNA genes in poorly studied phyla that were represented by only a handful of genomes in our complete genome dataset, such as candidate phyla Acetothermia (1 genome and 64 long-read sequences) and Patescibacteria (3 genomes and 330 long-read sequences).

Using the long-read dataset, we identified 18 additional phyla where unlinked rRNA genes are common, including several candidate phyla (BRC1, GAL15, WS1, WS2) and members of the Candidate Phyla Radiation (Patescibacteria, Figure 2). We also found several clades with high proportions of unlinked rRNA genes that had no representation in our complete genome dataset, including Rikenellaceae RC9 gut group (334/624), Verrucomicrobia genus *Candidatus Udaeobacter* (80/80), Atribacteria order Caldatribacteriales (37/37), Cyanobacteria order Obscuribacterales (4/4), Acidobacteria

Subgroup 2 (27/27), Planctomycetes order MSBL9 (40/40), and Chloroflexi class GIF9 (7/7). Overall, we found that 52% of the phyla covered in our combined datasets (37/71) have at least one representative with unlinked rRNA genes.

Unlinked rRNA genes are strongly conserved

We found that taxa with unlinked rRNA genes are not randomly distributed across bacterial and archaeal lineages - rather, we observed a strong phylogenetic signal for this trait, which we confirmed by calculating Pagel's lambda (lambda = 0.96, $p > 0.001$). To highlight this point, we assembled a phylogenetic tree from full-length 16S rRNA gene sequences representing both the complete genome dataset and the long-read metagenomic dataset. We found clusters of related taxa with exclusively unlinked rRNA genes (Figure 4) including: Euryarchaeota class Thermoplasmata, the vast majority of Deinococcus-Thermus, CPR division Patescibacteria, Verrucomicrobia DA101 group, Chloroflexi class Dehalococcoidia, and Alphaproteobacteria class Rickettsiales.

Genomic attributes associated with unlinked rRNA genes

Given that there are numerous bacterial and archaeal lineages where unlinked rRNA genes are commonly observed, we next sought to determine what other genomic features may be associated with this non-standard rRNA gene arrangement. We treated the presence of unlinked rRNA genes as a binary trait - if a genome had at least one unlinked rRNA gene we counted the genome as "unlinked". In our NCBI complete genome dataset, we found rRNA gene status to be conserved strongly at the species level - meaning that the majority of species had either exclusively linked or unlinked rRNA genes among their members (Supplemental Figure S2). Therefore, for the following tests, we used a subset of our NCBI complete genome dataset - retaining only a single representative of each species, unless the species had heterogeneous rRNA gene status (0.71% of species), in which case we retained one genome of each rRNA gene status. The analyses were corrected in order to account for the effect of phylogenetic structure in the data (see Methods).

Historically, unlinked rRNA genes have been strongly associated with the reduced genomes of obligate intracellular bacteria, implying that this trait may merely be a side effect of the strong genetic drift and weak selection these taxa experience (10-12). To test

this hypothesis, we compared the genome sizes of species with linked and unlinked rRNA genes using Phylogenetic Generalized Linear Models (phyloglm). While we found that genomes with unlinked rRNA genes had smaller genomes on average, this difference was not significant (Figure 5, phyloglm $p = 0.12$, means of groups: 4.15 Mbp linked, 2.72 Mbp unlinked).

The organization of rRNA genes within the same operon facilitates their joint regulation and co-expression at precise stoichiometric ratios. Selection for this trait is expected to be stronger in faster growing Bacteria and Archaea, where, at maximum growth rates, synthesis of the ribosome is the cell's chief energy expenditure (4). To test this hypothesis, we analyzed the association between the linkage of rRNA genes and traits related to rapid growth in Bacteria and Archaea. On average, genomes with unlinked rRNA genes had significantly fewer rRNA copies (Figure 5, phyloglm $p < 0.0001$, means of groups: 4.25 copies linked, 2.72 copies unlinked). We also calculated ΔENC' for each complete genome - a measure of codon usage bias that is negatively correlated with minimum generation time in Bacteria and Archaea (35). Interestingly, genomes with unlinked rRNA genes were predicted to have significantly longer minimal generation times (Figure 5, phyloglm $p=0.028$, means of groups: 0.23 linked, 0.18 unlinked). Additionally, in our long-read dataset we found that unlinked rRNA genes were more common in environments typified by slow growth rates; soil and sediment samples had higher proportions of unlinked rRNA genes than samples from anaerobic digesters and the human gut (Figure 3).

RNaseIII separates the precursors of the 16S and 23S rRNA from their common transcript for subsequent maturation and inclusion in the ribosome (2). RNaseIII is not an essential protein in most Bacteria and Archaea, and several phyla in which unlinked rRNA genes are common do not encode RNaseIII (e.g. Deinococcus-Thermus and Euryarchaeota; 41). Therefore, we checked if there was a significant association between unlinked rRNA genes and the presence of RNaseIII genes. Interestingly, we found that genomes with unlinked rRNA genes were significantly less likely to encode the bacterial form of RNaseIII genes (Figure 5 and Supplemental Figure S3, PF00636: phyloglm $p < 0.001$, means of groups: 1.0 linked, 0.71 unlinked; PF14622: phyloglm $p = 0.007$, means of groups: 0.86 linked, 0.66 unlinked). We were unable to check this relationship for archaeal RNaseIII, due

to the size of our archaeal dataset (phyloglm failed to converge, only 39 genomes in our dataset had this gene). However, we note that the archaeal RNaseIII PF11469 was found in only two clades that feature exclusively linked rRNA genes (Euryarcheaota family Thermococcaceae and Crenarchaeota family Thermofilaceae).

Discussion

While unlinked rRNA genes have been documented previously, we have demonstrated that they are far more widespread among Bacteria and Archaea than expected. We found that unlinked rRNA genes consistently occur in 12 phyla using a dataset of complete genomes (Figure 2C), and 18 additional phyla using a dataset of long-read metagenomic sequences obtained from environmental samples (Figure 2D). Interestingly, some phyla were classified as exclusively linked in our complete genome dataset, yet had many members with unlinked rRNA genes in our long-read dataset. For example, while there were no complete genomes in the phylum Verrucomicrobia with unlinked rRNA genes (0/32), 38% of verrucomicrobial rRNA sequences were unlinked in our long-read dataset (82/217), with the majority of this group closely related to the bacterium *Ca*. *Udaeobacter copiosus* from the DA101 soil group (42). This imbalance is likely due to the strong bias towards faster-growing organisms when using traditional cultivation methods (43), and the fact that cultivated Bacteria and Archaea still make up the majority of high-quality genomes in public databases (13). Our results highlight the importance of using a combination of complete genomes, where genetic organization and traits can be assessed rigorously, with metagenomic data that allows us to sample the diversity found in selected environments in an unbiased manner. Together, these independent datasets show that unlinked rRNA genes occur across many bacterial and archaeal phyla.

The widespread prevalence of unlinked rRNA genes in many environmental samples has important implications for the use of community analysis methods that require the 16S and 23S rRNA genes to be in close proximity. For instance, before 16S rRNA gene sequencing became common practice, the ITS region of the 16S and 23S rRNA operon was routinely used to fingerprint microbial communities (44). Likewise, the increasing popularity of long-read sequencing technologies has led to bacterial genotyping methods

that target the full rRNA operon. While sequencing from the 16S rRNA gene into the 23S rRNA gene (thus including the ITS region of the rRNA operon) can increase taxonomic resolution and allow strain level identification (45), our work shows that amplicon-based studies dependent on 16S and 23S rRNA genes being located in close proximity may miss a large portion of bacterial and archaeal diversity. We found the average distance between unlinked 16S and 23S rRNA genes in our complete genome dataset to be ~410 Kbp, a rather impractical distance to amplify by PCR. While strategies which use reads spanning the 16S and 23S rRNA genes to improve taxonomic resolution (e.g. 45,46) are less likely to introduce biases in some environments (e.g. human gut), they will miss many phylogenetic groups in other environments like soil and sediment, where a significant fraction of taxa have unlinked rRNA genes (Figure 3).

We used our long-read metagenomic dataset to not only bypass the cultivation bias of our complete genome dataset, but to also estimate the abundance of unlinked rRNA genes in a range of microbial community types. Our analyses of the long-read metagenomic dataset show that taxa with unlinked rRNA genes are far more abundant in some environments than others. Most notably, unlinked rRNA genes were far more common in soil (where as many as 41% of rRNA genes detected were unlinked) than the human gut (where no unlinked rRNA genes were detected, Figure 3). The environments with higher proportions of unlinked rRNA genes (soil and sediment) are generally thought to be populated by slower growing taxa (35,47). Likewise, we found that genomes with unlinked rRNA genes have significantly fewer rRNA copies than genomes with exclusively linked rRNA genes, a trait which is correlated with maximum potential growth rate (35,48). We also found that genomes with unlinked rRNA genes are predicted to have significantly longer generation times (using codon usage bias in ribosomal proteins as a proxy for maximal growth rates) compared to genomes with exclusively linked rRNA genes. These lines of evidence suggest that unlinked rRNA genes are more common in the genomes of taxa with slower potential growth rates.

The existence of numerous genomes that have unlinked 16S and 23S rRNA genes and the differential frequency of these genomes across environments raise the question of the role and implications of this genetic organization. Upon first consideration, having unlinked 16S and 23S rRNA genes would seem to be disadvantageous given that both rRNA

407 molecules are needed in equal proportions to yield a functioning ribosome. The importance
408 of linkage for identical expression of both rRNA genes should be greater in faster growing
409 taxa, where a higher rate of ribosome synthesis is key to rapid growth and accounts for a
410 large proportion of the cell energy budget (4). Studies in the fast-growing species *E.coli*
411 have shown that, while unbalanced rRNA gene dosage has a slight negative effect on
412 doubling times, balanced synthesis of ribosomal proteins still occurs in most cases (49). If
413 unequal expression of rRNA subunits is associated with unlinked rRNA genes, it may not
414 confer a selective disadvantage in many environments (like soils and sediments) where
415 longer generation times are the norm, not the exception. For slower-growing taxa, the
416 selection coefficient associated with the effect of linked rRNA genes on growth may be
417 small, because rRNAs are less expressed and rapid growth is a trait under weaker selection.
418 Under these circumstances, unlinked rRNA genes may become fixed in populations by
419 genetic drift. This is more likely to occur in species with small effective population sizes, i.e.
420 few effectively reproducing individuals, where natural selection is not efficient enough to
421 avoid the loss of genes or the degradation of genome organizational traits that are under
422 weak selection (50). This is the most common explanation for the occurrence of unlinked
423 16S and 23S rRNA genes (10-12). It fits our observations that many of the taxa we
424 identified with unlinked rRNA genes are restricted to obligate intracellular lifestyles
425 (including members of the phyla Spirochaetes, Epsilonbacteraeota, Alphaproteobacteria,
426 and Tenericutes) or contain signatures of symbiotic lifestyles (CPR phyla; 51,52).
427 However, fixation of mutations due to genetic drift is much less likely to explain the
428 presence of unlinked rRNA genes among the large proportion of free-living taxa that we
429 have identified (including members of the phyla Deinococcus-Thermus, Euryarchaeota,
430 Chloroflexi, Planctomycetes, and Verrucomicrobia). Some of these taxa are abundant and
431 ubiquitous in their respective environments, e.g. the Verrucomicrobia *Ca. U. copiosus* (42)
432 and members of the Rikenellaceae RC9 gut group (53). These genomes do not show traits
433 typically associated with genome reduction caused by small effective population sizes, i.e.
434 abundant pseudogenes, transposable elements, or small genomes. While we found that, on
435 average, the genomes of taxa with unlinked rRNA genes were smaller than those with
436 linked rRNA genes, this difference was not significant after accounting for phylogeny. Thus,
437 there is little evidence that the highly conserved trait of unlinked rRNA genes is caused

exclusively by genetic drift - especially in free-living taxa.

Unlinked rRNA genes could provide a selective advantage in certain circumstances, which may explain their existence in free-living taxa. Transcribing the 16S and 23S rRNA genes separately may eliminate or reduce the need for RNaseIII, which we found to occur in lower frequencies in taxa with unlinked rRNA genes (Supplemental Figure S3). We also found RNaseIII to be completely absent in the phyla Deinococcus-Thermus and Gemmatimonadetes, both phyla with high proportions of unlinked rRNA genes. Interestingly, two recent studies have investigated the function of RNaseIII in *Borrelia burgdorferi* (54) and *Helicobacter pylori* (55), two intracellular bacteria with exclusively unlinked rRNA genes. When RNaseIII was knocked out both bacteria remained viable, but accumulated unprocessed rRNA intermediates and exhibited decreased growth rates (54,55). On the other hand, some bacteriophages hijack host RNaseIII to process their own mRNA (56) - in some cases, host RNaseIII can stimulate the translation of infecting phage mRNA by several orders of magnitude (57) (although other phage appear indifferent to the presence of RNaseIII; 58). Regardless, increased resistance to predation at the cost of reduced maximum potential growth rates is a widely observed ecological trade-off (59). Lastly, recent work has shown that some rRNA loci specialize in the translation of genes involved in adaption to temperature and nutrient shifts (60). It is thus tempting to speculate that unlinked rRNA genes could facilitate the production of heterogeneous ribosomes with a diverse range of characteristics.

Conclusions

Unlinked rRNA genes are far more prevalent than expected, especially among those Bacteria and Archaea found in environmental samples for which complete genomes are not yet available. While this rearrangement appears to occur more frequently in slower-growing taxa and may be related to the presence of RNaseIII, it remains to be determined if unlinked rRNA genes confer any specific advantages. Regardless, we have shown that 52% of the phyla included in our combined datasets (37/71) have at least one member with unlinked rRNA genes, that unlinked rRNA genes occur in taxa that are abundant and ubiquitous, and that up to 41% of rRNA genes in some environments are unlinked - meaning unlinked rRNA genes are far from atypical anomalies. Indeed, unlinked rRNA

genes function as a reminder that the metabolisms of poorly-studied environmental Bacteria and Archaea sometimes differ from conventions derived from model organisms. We have developed hypotheses about the potential advantages of unlinked rRNA genes, hypotheses which could be tested experimentally and represent a promising direction for future research - especially as some taxa with unlinked rRNA genes are relatively easy to manipulate in culture (61,62).

Acknowledgements

This research was supported in part by the Chateaubriand Fellowship awarded to T.E.B. from the Office for Science & Technology of the Embassy of France in the United States and a grant to N.F. from the U.S. National Science Foundation (EAR1331828). M.A. was supported by a research grant (15510) from Villum Fonden. A.E. gratefully acknowledges the support of a Leverhulme Trust Research Fellowship (RF-2017-652\2). E.R. was supported by the INCEPTION project (PIA/ANR-16-CONV-0005). We thank Will Trimble for assistance tracking down publicly available Moleculo sequences, Michael Engel for figure design input, and Eric Johnston for introducing the lead author to unlinked rRNA genes.

Author contributions

TEB, ER, and NF conceived and designed the project and wrote the paper with input from all co-authors. AE, MA, and RK performed the Nanopore sequencing. TEB performed all analyses.

Conflict of interest statement

MA and RK own a portion of the company DNASense.

Data availability

All genomes used in this study were downloaded from NCBI, with assembly IDs listed in Supplemental Dataset S1. All Nanopore data is available at the Sequence Read Archive (SRA) under Bioproject ID PRJNA553237 or the European Nucleotide Archive (ENA) under PRJEB33278. All Moleculo data has been published previously, with

publications listed in methods. Classifications and details of both the complete genome and long-read datasets are included in Supplemental Dataset S1 and S2, respectively.

References

1. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. Nat Rev Micro. 2008 Mar 4;6:315–9.

2. Srivastava AK, Schlessinger D. Mechanism and Regulation of Bacterial Ribosomal RNA Processing. Annu Rev Microbiol. 1990;44:105–29.

3. Condon C, Squires C, Squires CL. Control of rRNA Transcription in *Escherichia coli*. Microbiological Reviews. 1995 Dec;59:623–45.

4. Gourse RL, Gaal T, Bartlett MS, Appleman JA, Ross W. rRNA Transcription and Growth Rate–Dependent Regulation of Ribosome Synthesis in *Escherichia coli*. Annu Rev Microbiol. 1996;50:645–77.

5. Klappenbach JA, Dunbar JM, Schmidt TM. rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. Applied and Environmental Microbiology. 2000 Apr;66:1328–33.

6. Hartmann RK, Ulbrich N, Erdmann VA. An unusual rRNA operon constellation: in *Thermus thermophilus* HB8 the 23S/5S rRNA operon is a separate entity from the 16S rRNA operon. Biochimie. 1987;69:1097–104.

7. Liesack W, Stackebrandt E. Evidence for Unlinked rrn Operons in the Planctomycete *Pirellula marina*. Journal of Bacteriology. 1989;171:5025–30.

8. Munson MA, Baumann L, Baumann P. *Buchnera aphidicola* (a prokaryotic endosymbiont of aphids) contains a putative 16S rRNA operon unlinked to the 23s rRNA-encoding gene: sequence determination, and promoter and terminator analysis. Gene. 1993;137:171–8.

9. Andersson SGE, Zomorodipour A, Winkler HH, Kurland CG. Unusual Organization of the rRNA Genes in *Rickettsia prowazekii*. Journal of Bacteriology. 1995;177:4171–5.

10. Rurangirwa FR, Brayton KA, McGuire TC, Knowles DP, Palmer GH. Conservation of the unique rickettsial rRNA gene arrangement in *Anaplasma*. International Journal of Systemic and Evolutionary Microbiology. 2002 Jul 1;52(4):1405–9.

11. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct. BioMed Central; 2009;4(13):13–25.

12. Andersson JO, Andersson SGE. Genome Degradation is an Ongoing Process in

Rickettsia. Molecular Biology and Evolution. 1999;16(9):1178–91.

13. Zhi X-Y, Zhao W, Li W-J, Zhao G-P. Prokaryotic systematics in the genomics era. Antonie van Leeuwenhoek. Springer Netherlands; 2012 Nov 25;101(1):21–34.

14. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016 Jan 3;44(D1):D733–45.

15. Team RC. R: A language and environment for statistical computing. 2018.

16. Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, et al. Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol. Nature Publishing Group; 2014 Feb 23;32(3):261–6.

17. Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat Biotechnol. Nature Publishing Group; 2016 Jan;34(1):64–9.

18. White RA, Bottos EM, Roy Chowdhury T, Zucker JD, Brislawn CJ, Nicora CD, et al. Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from Complex Soil Metagenomes. mSystems. American Society for Microbiology Journals; 2016 Jun;1(3):309–15.

19. Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, et al. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. Cold Spring Harbor Lab; 2015 Apr;25(4):534–43.

20. Flynn TM, Koval JC, Greenwald SM, Owens SM, Kemner KM, Antonopoulos DA. Parallelized, Aerobic, Single Carbon-Source Enrichments from Different Natural Environments Contain Divergent Microbial Communities. Front Microbiol. Frontiers; 2017 Nov 28;8:1540–14.

21. Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, et al. metaxa2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. Mol Ecol Resour. 2015 Mar 23;15(6):1403–14.

22. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology. 2007 Aug 10;73(16):5261–7.

23. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research. Oxford University Press; 2012;41(D1):D590–6.

24. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015 Jun

15;523(7559):208–11.

25. Pei A, Nossa CW, Chokshi P, Blaser MJ, Yang L, Rosmarin DM, et al. Diversity of 23S rRNA Genes within Individual Prokaryotic Genomes. PLoS ONE. Public Library of Science; 2009 May 5;4(5):1–9.

26. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Aug 12;26(19):2460–1.

27. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 2010 Jan 11;26(2):266–7.

28. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution. 2009 Jun 9;26(7):1641–50.

29. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research. 2016 Jul 4;44(W1):W242–5.

30. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999 Oct 20;401:877–84.

31. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. John Wiley & Sons, Ltd (10.1111); 2011 Dec 15;3(2):217–23.

32. Tung Ho LS, Ané C. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. Systematic Biology. 2014 Feb 4;63(3):397–408.

33. Novembre JA. Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias. Molecular Biology and Evolution. 2002 Jul 18;19(8):1390–4.

34. Rocha E. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 2004 Oct 17;14:2279–86.

35. Vieira-Silva S, Rocha E. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. PLOS Genetics. 2009 Dec 22;6(1):1–15.

36. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011 Oct 20;7(10):e1002195–16.

37. Moreno-Hagelsieb G, Collado-Vides J. A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics. 2002;18:S329–36.

38. Shepherd J, Ibba M. Bacterial transfer RNAs. FEMS Microbiol Rev. 2015 Mar

20;39(3):280–300.

39. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. Nature Publishing Group; 2018 Aug 27;36(10):996–1004.

40. Yuan C, Lei J, Cole J, Sun Y. Reconstructing 16S rRNA genes in metagenomic data. Bioinformatics. 2015 Jun 13;31(12):i35–i43.

41. Durand S, Gilet L, Condon C. The Essential Function of B. subtilis RNase III Is to Silence Foreign Toxin Genes. PLOS Genetics. 2012 Dec 27;8(12):e1003181–11.

42. Brewer TE, Handley KM, Carini P, Gilbert JA, Fierer N. Genome reduction in an abundant and ubiquitous soil bacterium "Candidatus Udaeobacter copiosus." Nature Microbiology. The Author(s) SN ; 2016 Oct 31;2:16198.

43. Vartoukian SR, Palmer RM, Wade WG. Strategies for culture of "unculturable" bacteria. FEMS Microbiology Letters. 2010 Apr 27;309:1–7.

44. Garcia-Martinez J, Acinas SG, Anton AI, Rodriguez-Valera F. Use of the 16S-23S ribosomal genes spacer region in studies of prokaryotic diversity. J Microbiolal Methods. 1999 Apr 22;36:55–64.

45. Zeng YH, Koblížek M, Li YX, Liu YP, Feng FY, Ji JD, et al. Long PCR-RFLP of 16S-ITS-23S rRNA genes: a high-resolution molecular tool for bacterial genotyping. J Appl Microbiol. John Wiley & Sons, Ltd (10.1111); 2012 Dec 20;114(2):433–47.

46. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon. F1000Res. 2018 Nov 6;7:1755–25.

47. Brown CT, Olm MR, Thomas BC, Banfield JF. Measurement of bacterial replication rates in microbial communities. Nat Biotechnol. Nature Publishing Group; 2016 Dec;34(12):1256–63.

48. Roller BRK, Stoddard SF, Schmidt TM. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nature Microbiology. 2016 Sep 12;1:1–7.

49. Siehnel RJ, Morgan EA. Unbalanced rRNA Gene Dosage and its Effects on rRNA and Ribosomal-Protein Synthesis. Journal of Bacteriology. 1985 Aug;163(2):476–86.

50. Moran NA. Microbial Minimalism: Genome Reduction in Bacterial Pathogens. Cell. 2002 Mar 1;108:583–6.

51. Nelson WC, Stegen JC. The reduced genomes of Parcubacteria (OD1) contain

signatures of a symbiotic lifestyle. Front Microbiol. 2015 Jul 21;6(110):693–14.

52. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nature Communications. Nature Publishing Group; 2016 Jan 27;7:1–8.

53. Holman DB, Brunelle BW, Trachsel J, Allen HK. Meta-analysis To Define a Core Microbiota in the Swine Gut. mSystems. 11 ed. 2017 May 23;2(3):676–14.

54. Anacker ML, Drecktrah D, LeCoultre RD, Lybecker M, Samuels DS. RNase III Processing of rRNA in the Lyme Disease Spirochete *Borrelia burgdorferi*. Journal of Bacteriology. American Society for Microbiology Journals; 2018 Jul 1;200(13):1–11.

55. Iost I, Chabas S, Darfeuille F. Maturation of atypical ribosomal RNA precursors in *Helicobacter pylori*. Nucleic Acids Research. Oxford University Press; 2019 Apr 22;47(11):5906–21.

56. Gone S, Alfonso-Prieto M, Paudyal S, Nicholson AW. Mechanism of Ribonuclease III Catalytic Regulation by Serine Phosphorylation. Nature. Nature Publishing Group; 2016 Apr 26;6(25448):1–9.

57. Wilcon HR, Yu D, Peters HK III, Zhou J-G, Court DL. The global regulator RNase III modulates translation repression by the transcription elongation factor N. The EMBO Journal. 2002;21:4154–61.

58. Hagen FS, Young ET. Effect of RNase III on Efficiency of Translation of Bacteriophage T7 Lysozyme mRNA. Journal of Virology. 1978 Feb 26;26:793–804.

59. Bohannan BJM, Lenski RE. Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. Ecology Letters. 2000 Jul 7;3:362–77.

60. Song W, Joo M, Yeom J-H, Shin E, Lee M, Choi H-K, et al. Divergent rRNAs as regulators of gene expression at the ribosome level. Nature Microbiology. Springer US; 2019;4:515–26.

61. Holland AD, Rothfuss HM, Lidstrom ME. Development of a defined medium supporting rapid growth for Deinococcus radiodurans and analysis of metabolic capacities. Appl Microbiol Biotechnol. 2006 Mar 31;72(5):1074–82.

62. Devos DP. Gemmata obscuriglobus. Current Biology. Elsevier; 2013 Sep 9;23(17):R705–7.

Figure Captions

Figure 1: In most Bacteria and Archaea, rRNA genes are arranged in the order 16S - 23S - 5S, and are transcribed and regulated as a single unit. However, in some cases, the 16S is separated from the 23S and 5S, and is referred to as “unlinked”.

Figure 2: Unlinked rRNA genes can be found in 30 phyla. A) The distribution of ITS lengths in complete genomes from NCBI. 1.3% of NCBI rRNA genes have an ITS region $>$ 1500 bp in length. The majority of unlinked rRNA genes have an ITS of $>$ 6000 bp (682/778) with a mean length of 410374 bp (±521792 bp). B) The distribution of ITS lengths in the long-read sequence dataset. 10.1% of rRNA genes have an ITS $>$ 1500 bp. The majority of unlinked genes have an ITS of unknown length due to sequence length constraints in the long-read dataset (1470/1607). C) Within our set of complete genomes from NCBI, 12 phyla had genomes containing at least one set of unlinked rRNA genes in $>$5% of members. Linked refers to genomes with exclusively linked rRNA genes, unlinked refers to genomes with exclusively unlinked rRNA genes, and mixed refers to genomes with at least one set each linked and unlinked rRNA genes. D) By analyzing long-read metagenomic datasets, we confirmed that 8 of the phyla with unlinked rRNA genes in the complete genome dataset also had unlinked rRNA genes in environmental samples (top portion), and found an additional 18 phyla in which $>$5% of reads that met our criteria for inclusion in downstream analyses (see Methods) contained unlinked rRNA genes.

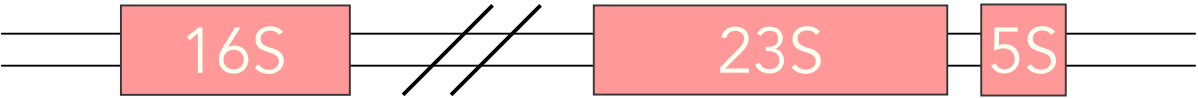
Figure 3: Unlinked rRNA genes have differential frequencies across environments. We found that soils (13-41% unlinked) and sediments (7.7-29%) have more unlinked rRNA genes on average than anaerobic digesters (8.1-8.8%) and the human gut (0%). Results obtained from analyses of Moleculo and Nanopore metagenomic data are indicated with (m) and (n), respectively.

Figure 4: Unlinked rRNA genes occur in coherent phylogenetic clusters. This phylogenetic tree was created from full-length 16S rRNA sequences by combining both the NCBI complete genome and long-read metagenomic datasets (details in Methods). The outer ring indicates which dataset each sequence originated from, while the inner ring indicates the

status of rRNA genes. Sequences originating from the long-read dataset cannot be mixed, as we could not distinguish multi-copy rRNA genes. Clades with high proportions of unlinked members *and* good representation in the tree are indicated in green: A) Euryarchaeota class Thermoplasmata, B) Spirochaetae classes Leptospirae and Spirochaetia, C) Patescibacteria, D) Chlorflexi class Dehalococcoidia, E) Planctomycetes classes Phycisphaerae and Planctomycetacia, F) Verrucomicrobia genus *Candidatus* Udaeobacter, G) Tenericutes genus *Mycoplasma*, H) Deinococcus-Thermus, I) Epsilonbacteraeota genera *Helicobacter* and *Campylobacter*, J) Alphaproteobacteria order Rickettsiales and K) Gammaproteobacteria genus *Buchnera*.

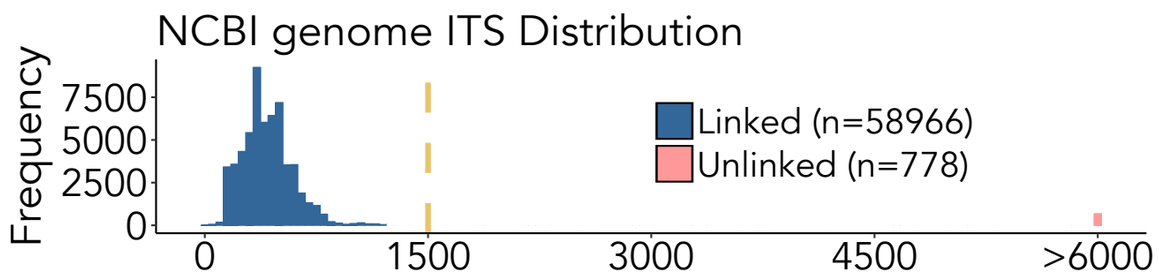
Figure 5: Genomic attributes of NCBI complete genomes based on their rRNA gene status. Linked genomes feature exclusively linked rRNA genes; unlinked genomes have at least one set of unlinked rRNA genes. We calculated these statistics using a subset of our complete genomes, including one genome per unique species and rRNA gene status. A) Genomes with unlinked rRNA genes have smaller genomes on average, but this difference was not significant after accounting for phylogeny (phyloglm $p = 0.12$, means of groups: 4.15 Mbp linked, 2.72 Mbp unlinked). B) On average, genomes with unlinked rRNA genes had significantly fewer rRNA copies (phyloglm $p < 0.0001$, means of groups: 4.25 copies linked, 2.72 copies unlinked). C) Genomes with unlinked rRNA genes are predicted to have longer average generation times (phyloglm $p = 0.028$, means of groups: 0.23 linked, 0.18 unlinked; as a reference E. *coli* has an average ΔENC' of 0.3). D) We found that there were significantly fewer RNaseIII genes in genomes with unlinked rRNA genes (only PF00636 shown, for more detail see Supplemental Figure S3: phyloglm $p < 0.001$, means of groups: 1.0 linked, 0.71 unlinked).

Canonical linked rRNA operon

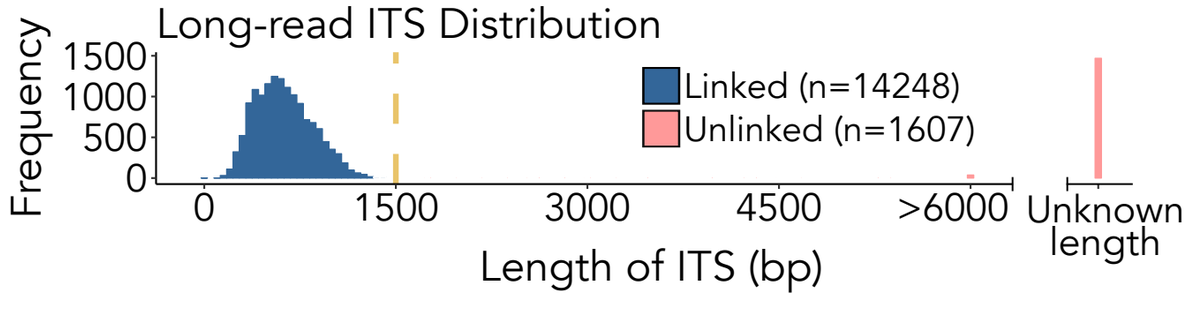


Unlinked rRNA genes

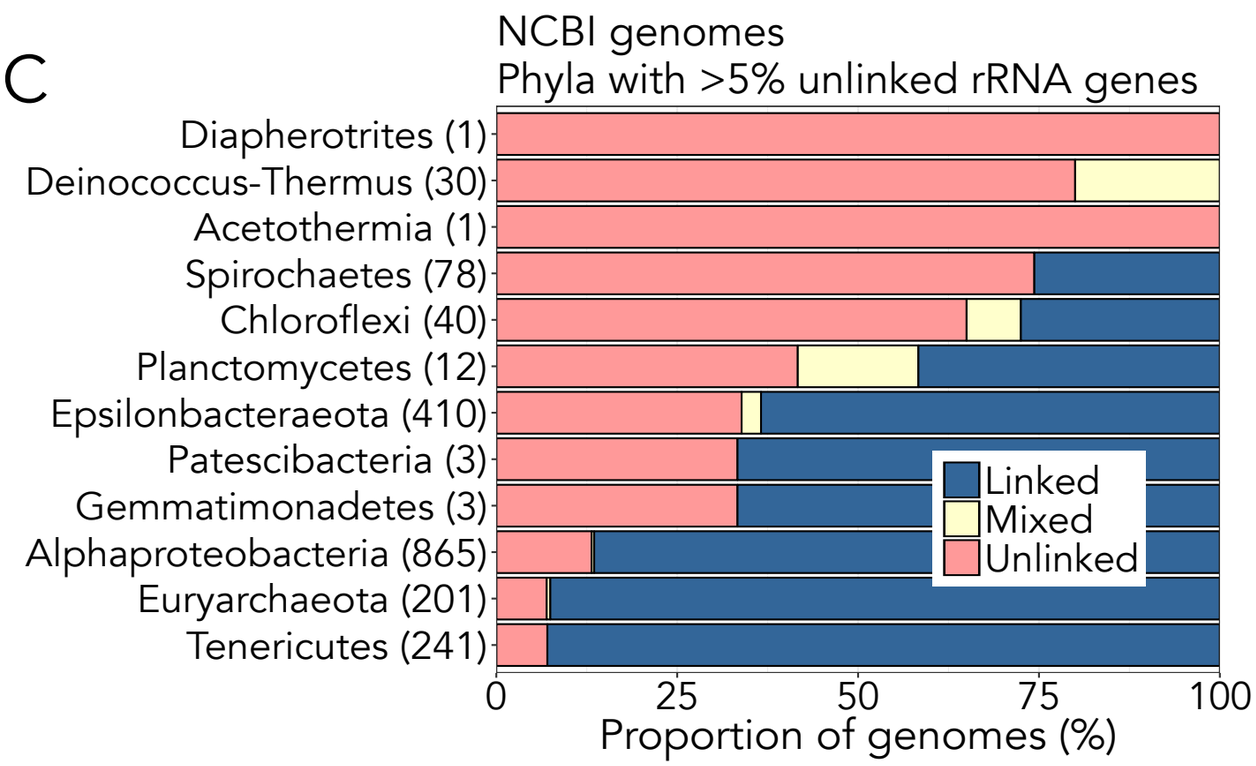
A


NCBI genome ITS Distribution
Frequency
7500
5000
2500
0
Linked (n=58966)
Unlinked (n=778)
0
1500
3000
4500
>6000

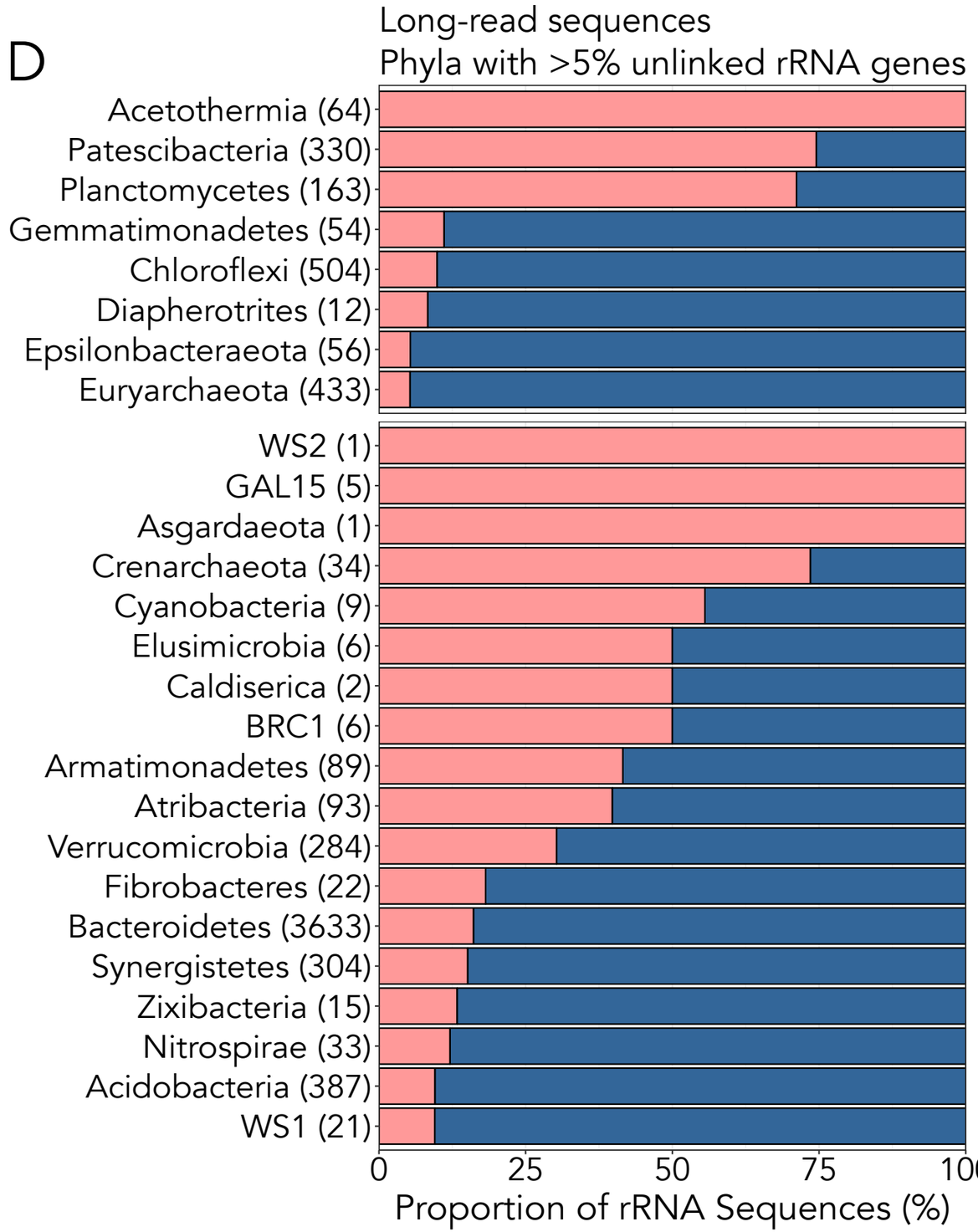
B

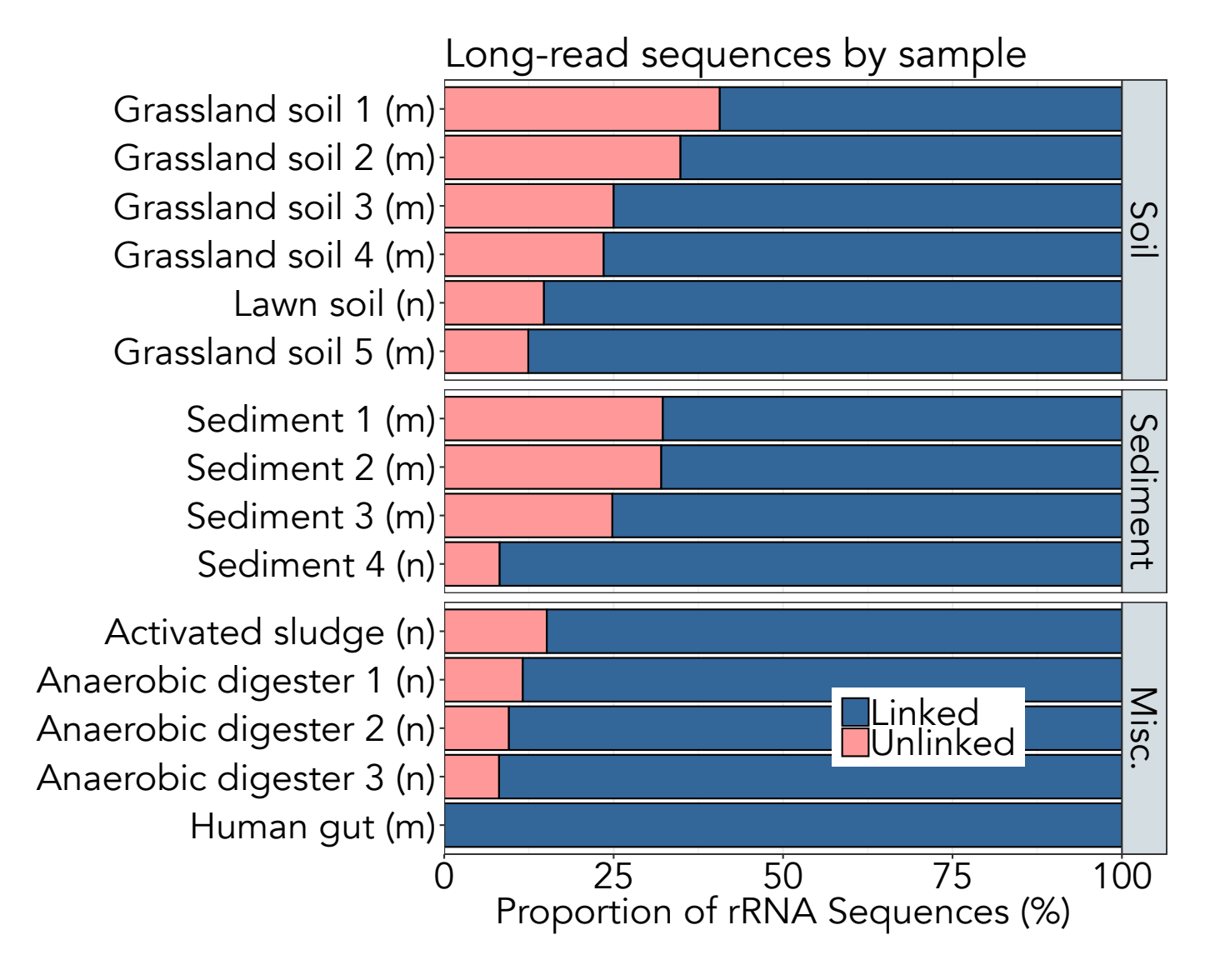

Long-read ITS Distribution
Frequency
1500
1000
500
0
Linked (n=14248)
Unlinked (n=1607)
0
1500
3000
4500
>6000
Unknown length
Length of ITS (bp)

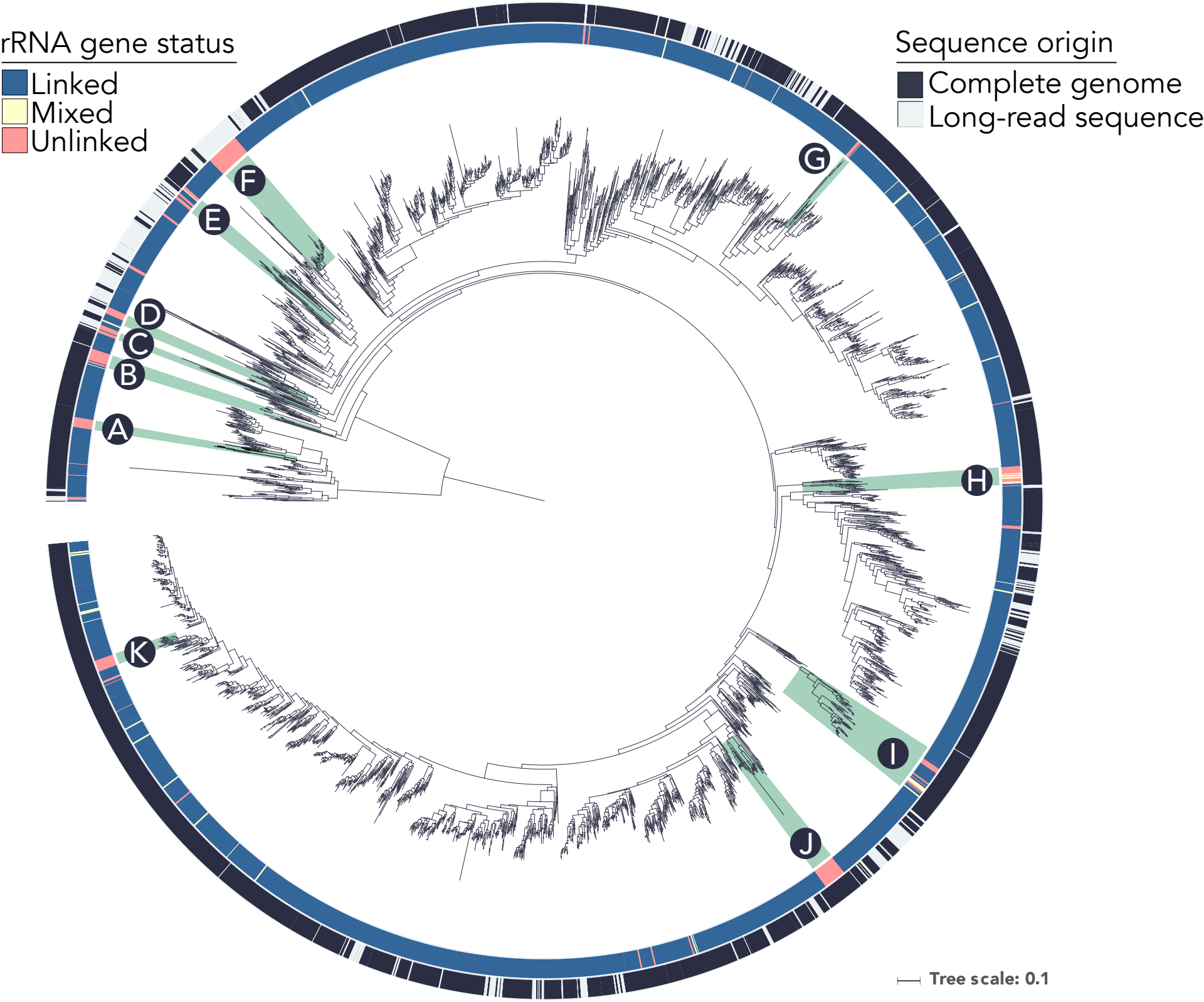
C

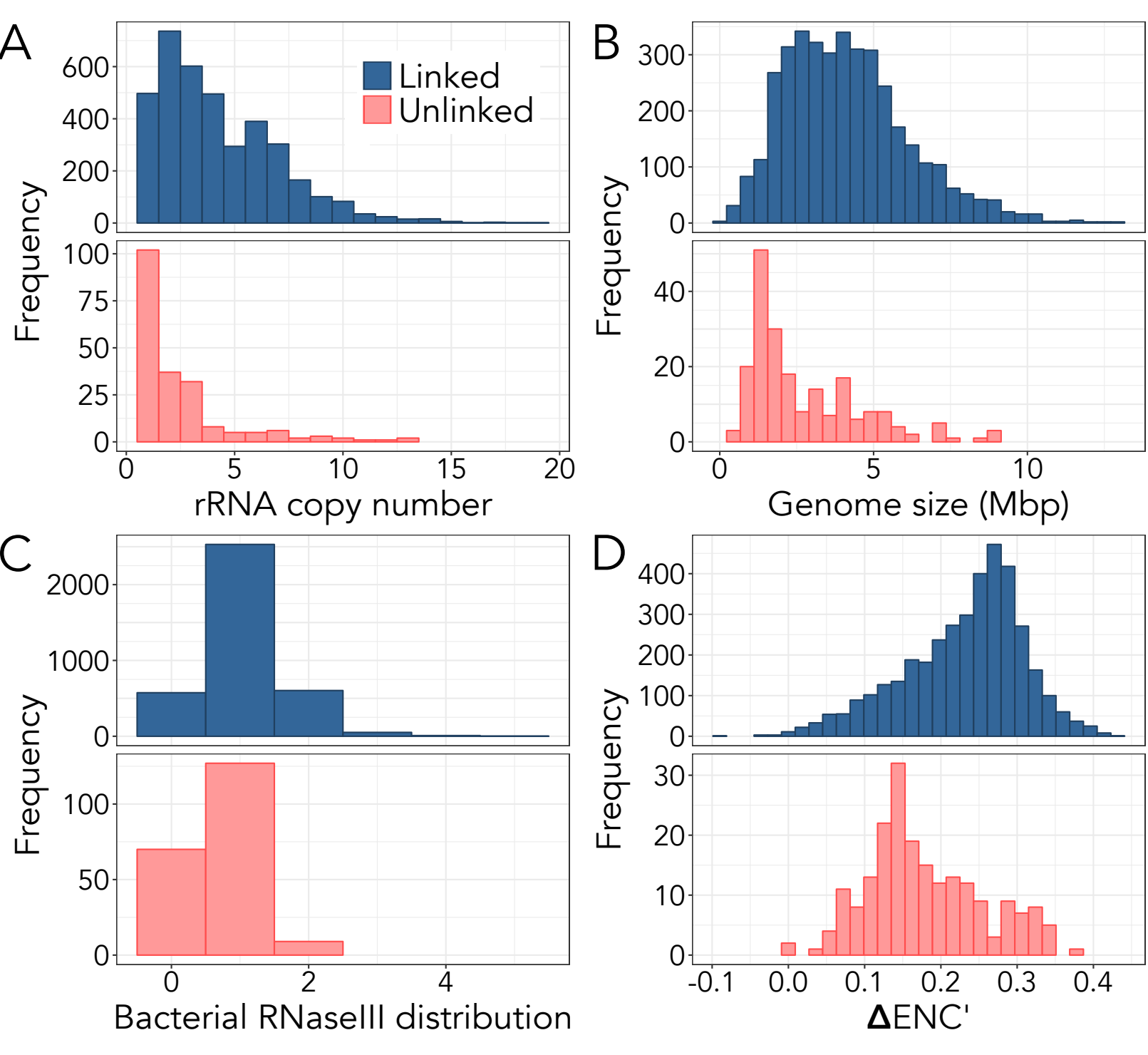

NCBI genomes
Phyla with >5% unlinked rRNA genes
Diapherotrites (1)
Deinococcus-Thermus (30)
Acetothermia (1)
Spirochaetes (78)
Chloroflexi (40)
Planctomycetes (12)
Epsilonbacteraeota (410)
Patescibacteria (3)
Gemmatimonadetes (3)
Alphaproteobacteria (865)
Euryarchaeota (201)
Tenericutes (241)
Linked
Mixed
Unlinked
0
25
50
75
100
Proportion of genomes (%)

D


Long-read sequences
Phyla with >5% unlinked rRNA genes
Acetothermia (64)
Patescibacteria (330)
Planctomycetes (163)
Gemmatimonadetes (54)
Chloroflexi (504)
Diapherotrites (12)
Epsilonbacteraeota (56)
Euryarchaeota (433)
WS2 (1)
GAL15 (5)
Asgardaeota (1)
Crenarchaeota (34)
Cyanobacteria (9)
Elusimicrobia (6)
Caldiserica (2)
BRC1 (6)
Armatimonadetes (89)
Atribacteria (93)
Verrucomicrobia (284)
Fibrobacteres (22)
Bacteroidetes (3633)
Synergistetes (304)
Zixibacteria (15)
Nitrospirae (33)
Acidobacteria (387)
WS1 (21)
0
25
50
75
100
Proportion of rRNA Sequences (%)


Long-read sequences by sample
Grassland soil 1 (m)
Grassland soil 2 (m)
Grassland soil 3 (m)
Grassland soil 4 (m)
Lawn soil (n)
Grassland soil 5 (m)
Sediment 1 (m)
Sediment 2 (m)
Sediment 3 (m)
Sediment 4 (n)
Activated sludge (n)
Anaerobic digester 1 (n)
Anaerobic digester 2 (n)
Anaerobic digester 3 (n)
Human gut (m)
Soil
Sediment
Misc.
Linked
Unlinked
0
25
50
75
100
Proportion of rRNA Sequences (%)


rRNA gene status
Linked
Mixed
Unlinked
Sequence origin
Complete genome
Long-read sequence
A
B
C
D
E
F
G
H
I
J
K
Tree scale: 0.1


A
Linked
Unlinked
Frequency
600
400
200
0
100
75
50
25
0
0
5
10
15
20
rRNA copy number
B
300
200
100
0
Frequency
40
20
0
0
5
10
Genome size (Mbp)
C
2000
1000
0
Frequency
100
50
0
0
2
4
Bacterial RNaseIII distribution
D
400
300
200
100
0
Frequency
30
20
10
0
-0.1
0.0
0.1
0.2
0.3
0.4
ΔENC'

sample	type	file_type	total_sequences	sequences > 1000	median_length
LIB-RHK-1851	nanopore	fastq	6194277	4953661	3255
20180216_SMK_J3	nanopore	fastq	3747204	3747204	6023
LIB-RHK-1848	nanopore	fasta	3362711	2653517	3228
JMJ	nanopore	fastq	2775301	2114004	4212
MHA-58	nanopore	fastq	1784659	1650522	5375
VCsoil	nanopore	fasta	1751625	1751625	2456
SRR3505613	moleculo	fastq	247328	247328	7197
SRR2822456	moleculo	fastq	130702	130702	7808
KA3UB14	moleculo	fasta	115256	93161	8850
SRR1605785_sedin	moleculo	fastq	95045	95045	7317
SRR1605725_sedin	moleculo	fastq	76499	76499	7863
SRR1605797_sedin	moleculo	fastq	73515	73515	7859
KA3FB3	moleculo	fasta	67177	60415	9774
KA3FB14	moleculo	fasta	50850	40895	4548
KA3UB3	moleculo	fasta	34170	28877	8527

total_lsu_hits	total ssu_hits	sequences_passing_filters	environment	sample_name_fig3
21463	17761	28056	Misc.	Anaerobic digester 3 (n)
6049	4906	7858	Sediment	Sediment 4 (n)
11955	9842	15672	Misc.	Anaerobic digester 2 (n)
6970	5777	9172	Misc.	Anaerobic digester 1 (n)
4273	3473	5577	Misc.	Activated sludge (n)
2658	1976	3817	Soil	Lawn soil (n)
248	213	328	Soil	Grassland soil 5 (m)
692	534	878	Misc.	Human gut (m)
229	213	367	Soil	Grassland soil 2 (m)
274	253	405	Sediment	Sediment 2 (m)
232	196	325	Sediment	Sediment 1 (m)
258	187	325	Sediment	Sediment 3 (m)
135	129	207	Soil	Grassland soil 1 (m)
88	57	124	Soil	Grassland soil 4 (m)
69	69	112	Soil	Grassland soil 3 (m)